The China Quarterly (2024), 1–15
doi:10.1017/S0305741024000948

RESEARCH REPORT

Mitigating Missingness in Analysing Chinese Policy and Implications for the Fragility of Our Knowledge Base

Vincent Brussee

Faculty of Humanities, Leiden University, Leiden, the Netherlands
Email: v.w.d.brussee@hum.leidenuniv.nl

Abstract

Access challenges for China researchers have increased, including for online research. This paper focuses on one subset of such challenges: policy documents. As no studies have to date analysed variation in data availability over time, researchers studying official documents risk conflating variation in transparency with actual policy change. This paper analyses missingness and finds that publication of policy documents under China's "open government information" initiative increased until the mid-late 2010s but then began to decrease. A key determinant of policy transparency is whether a document is related to citizens' daily lives, as opposed to national security. Furthermore, nearly 20 per cent of policy documents become unavailable two years after their publication. The paper concludes with a discussion on how to mitigate these challenges.

摘要

中国学者面对的研究权限挑战日益恶化，包括在线研究。 本文章重点关注此类挑战的其中一个问题：政策文件。 因为至今还没有分析政策文件在不同时间下数据可用性的研究，所以研究官方文件的中国学者可能会将透明度的变化与实际政策的变化混为一谈。 本文分析数据缺失，显示在政府信息公开倡议下，在2010年代中后期前，政策文件发布量一直增加，但此后有所减少。决定透明度的关键因素是文件是否与公民日常生活还是与国家安全有具体的关系。此外，近百分之二十的政策文件在发布两年后就无法获得。 本文最后讨论了如何化解这些挑战。

Keywords: China; policy; transparency; fragility; methodology; missingness

关键词: 中国; 政策研究; 政务公开; 缺失数据; 方法学

The "fragility" of our knowledge base has resurfaced as a profound concern in China Studies.[1] Restrictions on on-the-ground research have been steadily increasing over the past years.[2] Online access has also witnessed increasing controls[3] and a decline in the publication of, for example, court verdicts.[4] Consequently, researchers and scholars are now confronted with a full spectrum of access challenges, which affects both research conducted within China and research conducted from afar using the internet. Restrictions now extend well beyond just a few sources dealing with sensitive topics such as human rights.

This paper focuses on one specific aspect of this "fragility": policy documents.[5] China's political system is "text-centred,"[6] and official documents are the tools that transform abstract ruling

[1] Tiffert 2019.
[2] Shambaugh 2024, 327–28.
[3] Brussee and Von Carnap 2024.
[4] Liebman et al. 2023.
[5] Shorthand for any regulatory or normative document (*guifanxing wenjian*).
[6] Van de Ven 1995.

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

ideology into daily politics.[7] Perhaps because of the relative ease of obtaining policy documents, there is a rapidly emerging field that researches policy change by analysing textual changes in official documents. In recent issues of *The China Quarterly*, for instance, Abbey Heffer and Gunter Schubert analyse the introductions to policy documents taken from PKULaw, the Peking University database of laws and policies (*Beida fabao* 北大法宝) to illustrate the increasing use of policy experimentation in contemporary China,[8] and Yuen Yuen Ang mines central documents to research changes in policy communication.[9] Scholars also quantitatively use policy documents to examine the implementation of policy in specific domains, such as aging policies[10] and the Belt and Road Initiative, among others.[11]

Regrettably, as this paper discusses in more detail below, these studies seldom discuss whether their findings could be affected by changes in the availability of data over time, as opposed to actual changes in policy. The urgency of mitigating missingness has already been demonstrated in other contexts related to official documents from China, such as court judgments,[12] but similar insights are non-existent in the field of policy.

This paper discusses how researchers can manage variation in data availability when analysing official documents from China. To illustrate the need to reflect on missingness, it first reviews how the existing literature uses policy documents as data. It then discusses how the implementation of China's Open Government Information (OGI) framework can affect availability of data, before explaining the methods it uses to identify and analyse variation. The paper continues by offering empirical evidence of three types of variation. It concludes with methodological strategies and best practices to mitigate these. Altogether, it presents a word of caution when using these data in studies of Chinese policy and politics.

The Importance of Reflecting on Missingness in Policy Studies

Superficially, the emergence of access restrictions in China may seem not to apply to the study of policy documents. PKULaw remains available, without major restrictions, although it requires a subscription to access its full contents and has taken extensive measures to prevent the scraping of content. Government websites, like the State Council's database of central and ministerial-level policies, are mostly available without restrictions, too. This can make policy databases a highly convenient dataset for many researchers.

Perhaps because of a lack of overt restrictions, academic use of policy documents is rarely accompanied by a discussion of data limitations. Few papers (albeit with exceptions) explicitly discuss whether observations could be influenced by the varying availability of data, as opposed to actual changes in the documents. For example, while Yan Nan and colleagues highlight that "the number of aging policies in China increased rapidly since 2000" and suggest that this reflects changing pressures on the government,[13] they do not discuss the possibility that the Chinese government has increased not the number of policy documents it formulates but only those it actually releases to the public.[14] Since China's OGI framework only began to take shape in the early 2000s and was formalized nationally in 2008, this is a realistic concern: there are currently nearly 15,000 official central-level documents available on PKULaw that were originally published in 2008, versus only 2,610 for 1990.

[7] Wu, Guoguang 1995.
[8] Heffer and Schubert 2023.
[9] Ang 2024.
[10] Nan et al. 2020.
[11] Alves and Lee 2022.
[12] See, e.g., Liebman et al. 2020; 2023.
[13] Nan et al. 2020, 10–11.
[14] Similar types of arguments are made in, e.g., Huang et al. 2020, 332; Zhang et al. 2018.

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

Although some researchers attempt to mitigate for variation in data availability by using normalized data in studying changes over time,[15] normalization as a sole mitigating approach requires that variation is randomly distributed and that there is no variation in transparency for the specific categories they measure. Unfortunately, these papers do not explicitly articulate these limitations nor do they discuss the motivations for their mitigating strategies.

This implicit assumption is a risky one, as observable patterns would dramatically change if authorities, from one day to the next, decided to improve or restrict publication of documents in these categories. While established statistical strategies exist for mitigating randomly distributed variation, this is not the case for non-random variation.[16] Yet, in the context of archival censorship, Glenn Tiffert finds that omitted articles were not distributed randomly.[17] In the study of court judgments, scholars have highlighted that missingness affects particular categories more than others.[18] Although no studies to date have reflected on this in the field of policy, authorities do formulate annual guidelines on what information should be prioritized for (non-)disclosure. In the case of policy experimentation, for instance, some State Council documents call for an increase in publicity of certain government pilots.[19] This indicates that transparency may fluctuate for this category, which can risk conflating such variation with actual policy change.

This point is not to imply that any current findings are invalid. Heffer and Schubert, for instance, triangulate their findings with qualitative case studies and interviews with officials. Ang notes explicitly that her work should be seen as a "pilot" for what might be possible with these data.[20] The point is that it is crucial to discuss missingness and develop best practices to mitigate it.

Understanding Variation through the Lens of Open Government Information

To understand how variation might occur, it is crucial to consider the context in which government documents are disseminated – the OGI framework. While myriad studies have focused on the OGI as an object of study, few have examined it in light of the opportunities for analysis granted by information published under the OGI. This section argues researchers need to carefully consider two factors. First, the framework of the OGI ensures that the availability of data has never been consistent. Second, in recent years, the central government has increasingly raised security concerns related to the OGI, which escalates the urgency of research into data missingness.

Formalized nationally in 2008, authorities principally regard the OGI as a means to an end.[21] This reflects how law in China remains narrowly purpose oriented.[22] Such aims include resolving principal-agent dilemmas in policy implementation,[23] fighting corruption[24] and informing citizens and businesses about the regulations they need to comply with. As a result, while an increasing number of regulations have institutionalized disclosure of information deemed essential to the greater public, which in many regards falls in line with international practices, there are a great number of broad exemptions from disclosure and little judicial elaboration on their meaning.[25]

The implication is that authorities may interpret vague guidelines or legal norms in accordance with their (shifting) priorities or institutional constraints. In the Chinese-language literature, some scholars have highlighted "diametrically opposed" administrative practices that are rooted in

[15] Heffer and Schubert 2023, 43; Ang 2024.

[16] For a discussion of the randomness problem, see Xi 2022.

[17] Tiffert 2019, 556–57.

[18] Liebman et al. 2023.

[19] General Office of the State Council 2017.

[20] Ang 2024, 35.

[21] Horsley 2007, 63; Stromseth, Malesky and Gueorguiev 2017, 17.

[22] deLisle 2017; Creemers 2021.

[23] Chen, Liu and Tang 2022, 730.

[24] Horsley, 2007, 63.

[25] Horsley 2019, 522.

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

different applications of exemptions related to "work secrets" and "internal affairs."[26] Others refer to the general inability and unwillingness of some government agencies to implement the OGI[27] and the persistence of large gaps between various departments.[28] In the context of this study, the result is a variation in data availability, which must be mitigated for.

Since the 2020s, the central authorities have increasingly expressed heightened security concerns, which may further compound the variation in data availability. Although concerns with the political risks related to the OGI are not new, two notable changes in policy discourse since the 2020s hint at changing priorities. First, recent documents are downplaying commitments to transparency. In 2023, the State Council amended its Work Regulations and removed two *tifa* 提法[29] that had featured in almost all high-level documents relating to the OGI since 2014: to make transparency a fundamental principle of government work and to make disclosure the norm.[30] Moreover, the regulations now emphasized the goals of and considerations for disclosure ("to disclose according to law") over disclosure as a principle for its own sake. This changes the nature of the effort. While "transparency" implies a higher principle, the new language emphasizes the instrumentalist nature of the OGI.

Second, documents are expressing heightened security concerns over disclosure. For instance, the emphasis of the 2022 version of the annual "Work priorities for open government affairs" was on improving the OGI confidentiality review system, strictly conducting confidentiality reviews, preventing leaks not just of state secrets but also of "sensitive information" and preventing risks caused by data aggregation.[31] This was the first time a State Council-level document has mentioned these types of risk. Furthermore, it demanded that authorities "comprehensively consider the purpose, effect, and subsequent impact of disclosure."[32] Finally, the document encouraged the development of "scientific and rational" ways to determine the scope of publication, clarifying that authorities should consider disclosing some information to selected stakeholders only.

It is impossible to say, at this time, how these changes in policy discourse will be interpreted and implemented by state agencies. However, when seen in the broader context of information sources disappearing, it appears highly unlikely that they will be completely ignored. In fact, concrete indicators of change can already be seen. Most strikingly, the State Council did not promulgate or publish the 2023 version of its "Work priorities for open government affairs." This is the first time since 2012 that has not done so. These annual publications are important calls to action and failure to publish them marks a significant departure in transparency practices. More indicators of this change are discussed below in the results section.

Methods

The remainder of this paper empirically identifies and discusses variation in policy transparency, drawing from earlier scholarly precedents in missingness analysis.[33] Between 2021 and 2023, custom-made web scrapers retrieved around 310,000 policy and policy-adjacent documents from over 80 official websites of national and provincial Party and state organs.[34] This section provides a brief overview of the methods used; a more detailed discussion can be found in the Appendix.[35]

[26] Hu 2023, 17–18.

[27] Zhou 2016, 5.

[28] Chen et al. 2023, 15.

[29] *Tifa* is a political catchphrase that is deliberately used to convey a political signal.

[30] State Council 2018, Art. 28; State Council 2023, Art. 16.

[31] General Office of the State Council 2022.

[32] Ibid., Art. 12.

[33] See, e.g., Tiffert 2019, 554–55; Wu, Xiaohan, et al. 2022.

[34] Policy-adjacent documents include speeches given by leadership figures, meeting records of government organs, policy interpretations, etc.

[35] The project code and subsets of data are available at: https://github.com/zongtihuoguoguan/Policy-Transparency-China-2024.

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

Variation across time

To analyse variation across time, this paper analyses the serial numbers (*fawen zihao* 发文字号) of Chinese policy documents.[36] For instance, we may have access to documents numbered 1–5 and 7–10, but number 6 could be unavailable to the public. I first applied the "German tank problem" to estimate the actual total number of documents (i.e. documents after the last known number), before mapping these numbers to find patterns in transparency over time.[37]

Nationally, this analysis covers the four principal types of state documents: the *guofa* 国发 and *guobanfa* 国办发, which represent the high-authority documents issued by the State Council and its General Office, and the *guohan* 国函 and *guobanhan* 国办函, which generally include organizational documents. Provincial documents mirror this structure, although not all consistently provide the document numbers of their policies. Hence, I only include provinces with representative data in this analysis.

Variation across policy types and content

To analyse variation across policy types and policy content, this paper analyses policy referrals. Policy documents in China regularly refer to other policies, either to signal alignment with higher-level directives or to indicate future policy releases, even if the higher-level directive has never been made public. Custom scripts parsed these titles from the dataset and cross-referenced them with all published official documents. After tokenizing the titles, the "Fightin' words" algorithm identified discriminating words for public and non-public documents.[38] Afterwards, I manually selected the most context-relevant terms for analysis.[39]

Although it is impossible to identify the actual date of publication for those documents of which we do not have a full text, I used the date a policy was first mentioned elsewhere as a proxy timestamp. I applied a dictionary method, whereby each "topic" is defined by a series of keywords, to map patterns over time for different topics.

Variation owing to (dis)appearing documents

To assess the severity of disappearing or deleted documents, another custom script randomly sampled 50 links for each of the source websites (over 4,000 links in total) and verified whether the full content was still available. The sample consists of documents from 2021 exclusively, as this was when scraping started and when data should not be affected by deletion. For each unavailable document, I used the Wayback Machine to determine whether the link was unavailable because of website updates or because the document was individually deleted from the website.

Alongside making documents disappear, authorities can also make documents appear by retroactively publishing documents. To assess this, I automatically calculated the number of days between the issuance of a document, which is when it is formalized but not necessarily released to the public, and its publication. As not all government agencies consistently display issuing dates vis-à-vis publishing dates, I again relied on a subsection of agencies with relatively complete data.

Results

This section discusses the results of the analysis for each type of variation in turn.

Policy transparency is in decline at the top, yet effects are not uniform

While transparency increased significantly in the early-to-mid 2010s, there have been significant steps backwards in more recent years. Figure 1 displays the transparency rates of State Council documents

[36] Derived from Batke, Breuer and Stepan 2016.

[37] Clark, Gonye and Miller 2021.

[38] Monroe, Colaresi and Quinn 2008.

[39] See Appendix.

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

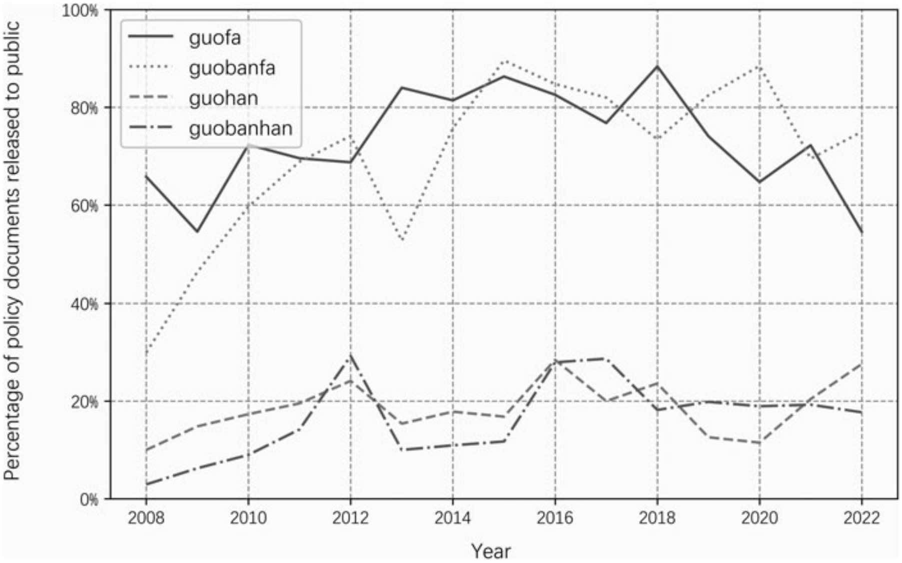

Figure 1. Transparency Rates of State Council Documents by Type of Policy

from 2008 to 2022. It shows that disclosure levels of the top-level *guofa* and *guobanfa* documents follow an inverted parabolic shape: increasing from 2008 to the mid-2010s but decreasing thereafter. In 2018, disclosure of *guofa* documents peaked at 88 per cent and then declined after, to 54.5 per cent in 2022. For *guobanfa* documents, disclosure decreased from 88 per cent in 2020 to 75 per cent in 2022. The disclosure of lower-level documents is consistently inconsistent (discussed below).

Figure 2 repeats this analysis for provincial documents and shows that availability there varies greatly, too. Some provinces consistently report high policy transparency rates; the rates for other provinces only started to climb in more recent years. Similarly, while some provinces issue some *han* 函 documents to the public, others do not. Furthermore, figures for some of the provinces assessed here show significant decreases in more recent years. In 2022, transparency figures for top-level documents from Henan (*yuzheng* 豫政), Shanghai (*huzhengfa* 沪政发), Hubei (*ezhengfa* 鄂政发) and Guangdong (*yuefu* 粤府) all dropped to their lowest levels in eight or more years. Nevertheless, these decreases remain minor in comparison with the increase in transparency since 2008.

The key determinant of transparency is a policy's relationship to citizens' daily lives

Another reason patterns are far from uniform is because of the variation between policy fields and types. Table 1 shows that regulations, plans and guiding documents are associated with disclosure. Meanwhile, internal policy processes such as reports, requests for approvals and evaluations are associated with non-disclosure. This aligns with relevant provisions that require the proactive disclosure of documents that are immediately relevant to citizens' daily lives but which also contain exceptions for internal processes.

This pattern continues, as shown in Table 2, which shows that topics closely related to people's daily lives are typically more transparent than those related to internal processes, security, the Party and strategy. Moreover, "science and technology," a topic closely related to ongoing US–China tensions, is also associated with non-disclosure. This is not an artefact in keyword selection: similar terms that are also associated with non-disclosure include "science" (*kexue* 科学: -2.3), "information technology" (*xinxi jishu* 信息技术: -2.1) and the Ministry of Science and Technology itself (*kexue jishu bu* 科学技术部: -2.3).

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

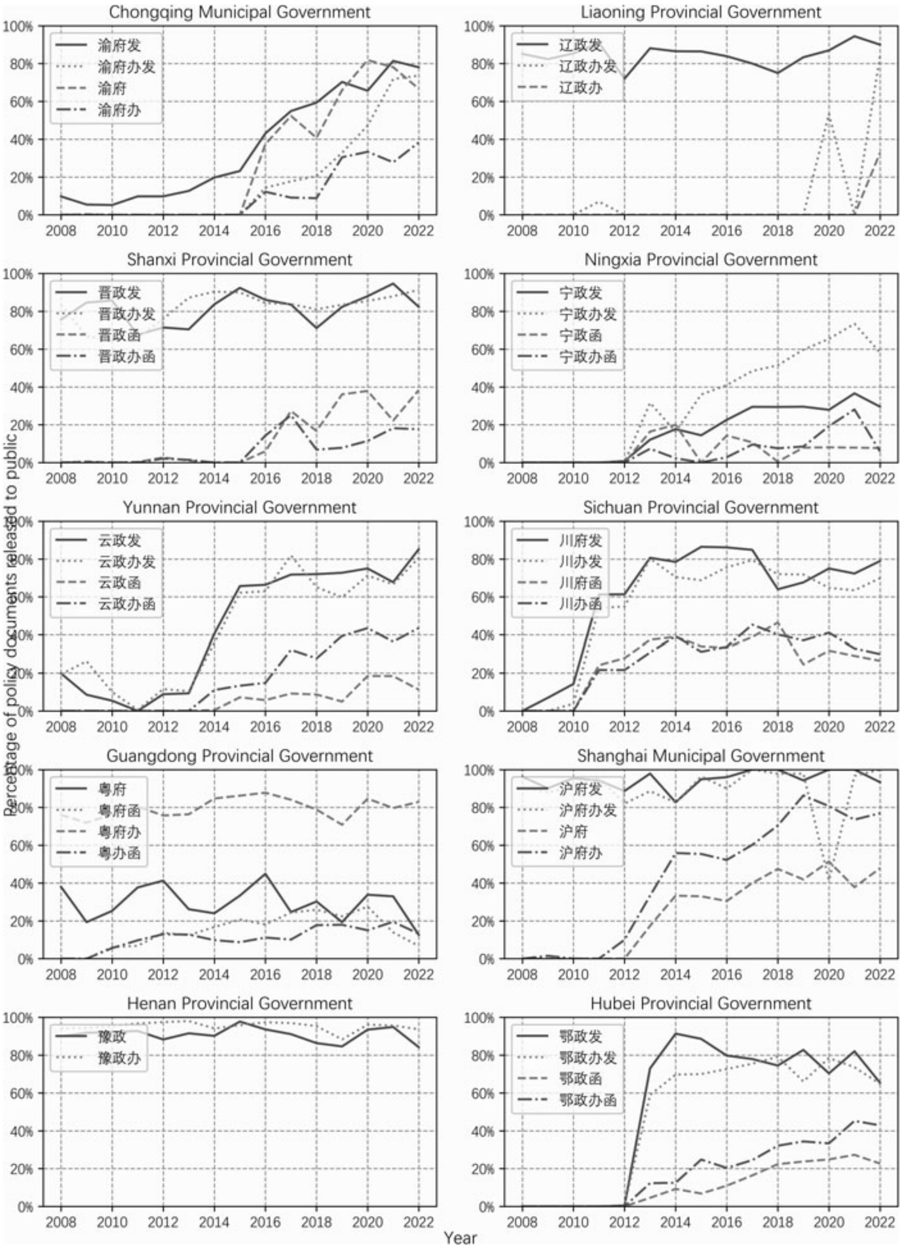

Figure 2. Transparency Rates of Provincial Policies by Type of Policy

The most high-profile example is the 14th Five-Year Plan on Science, Technology and Innovation, which has not been released to the public. However, local policy documents since 2021 confirm its existence.[40] The ongoing technological competition between the US and China is a key driver behind keeping this document out of the public domain. The plan covers many technologies that are subject to geopolitical competition. Furthermore, earlier strategy documents in this

[40] People's Government of Jiangsu Province 2021.

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

Table 1. Distinctive Word Analysis, Keywords Referring to Policy Types

Strongest Association with Disclosure			Strongest Association with Non-disclosure		
Chinese	English	Score	Chinese	English	Score
yijian 意见	opinions	23.78	*baogao* 报告	report	–17.56
banfa 办法	measures	17.73	*qingkuang* 情况	situation	–14.58
jueding 决定	decision	13.22	*qingshi* 请示	request for approval	–12.47
guihua 规划	plan	12.51	*pingjia* 评价	evaluation	–9.84
guiding 规定	provisions	12.44	*yaoqiu* 要求	requirements	–9.79
gangyao 纲要	outline	7.00	*ti'an* 提案	proposal	–4.90

Table 2. Distinctive Word Analysis, Keywords Referring to Policy Content

Strongest Association with Disclosure			Strongest Association with Non-disclosure		
Chinese	English	Score	Chinese	English	Score
jiaotong yunyu 交通运输	transport	13.93	*yanjiu* 研究	research	–11.97
jiandu guanli 监督管理	(market) supervision	8.37	*zhaopai* 招聘	recruitment	–6.20
nongye 农业	agriculture	7.60	*shuju* 数据	data	–6.24
fangkong 防控	(pandemic) prevention and control	7.15	*ganbu* 干部	cadre	–5.94
xiaofei 消费	consumption	6.46	*kexue jishu* 科学技术	science and technology	–5.67
yanglao 养老	pensions	6.05	*xitong* 系统	systems	–5.53
zhuanxiang 专项	special projects	5.91	*diaocha* 调查	investigations	–5.02
fuwu 服务	services	5.75	*renmin jingcha* 人民警察	People's Police	–3.83
yiliao weisheng 医药卫生	medicine	5.56	*zhonggong* 中共	the Party	–3.82
jijin 基金	funds	4.68	*shilian* 试验	experimentation	–3.80
yiliao baozhang 医疗保障	medical insurance	4.53	*zhanlüe* 战略	strategy	–3.71

field, such as the Made in China 2025 plan, triggered concerns in advanced economies about China's technical capabilities.[41]

In addition to this static picture, Figure 3 presents the transparency levels of different topics over time and demonstrates that the static patterns also hold over time. Topics closely related to people's lives (for example, the environment, education, socioeconomic policy) have witnessed increasing policy transparency rates; however, topics further removed from daily life (for instance, science and technology, state-owned resources, cadre management, international affairs) have decreased since 2014.[42]

The (dis)appearance of policy documents creates variation in document availability

Figure 4 shows the availability of links to policy documents two years after the date they were originally retrieved. Only 80.2 per cent of links were still available two years later; a further 10 per cent

[41] Cyrill 2018.

[42] Note that because of differing methodological approaches, individual figures are not comparable between figures 1–2 and 3.

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

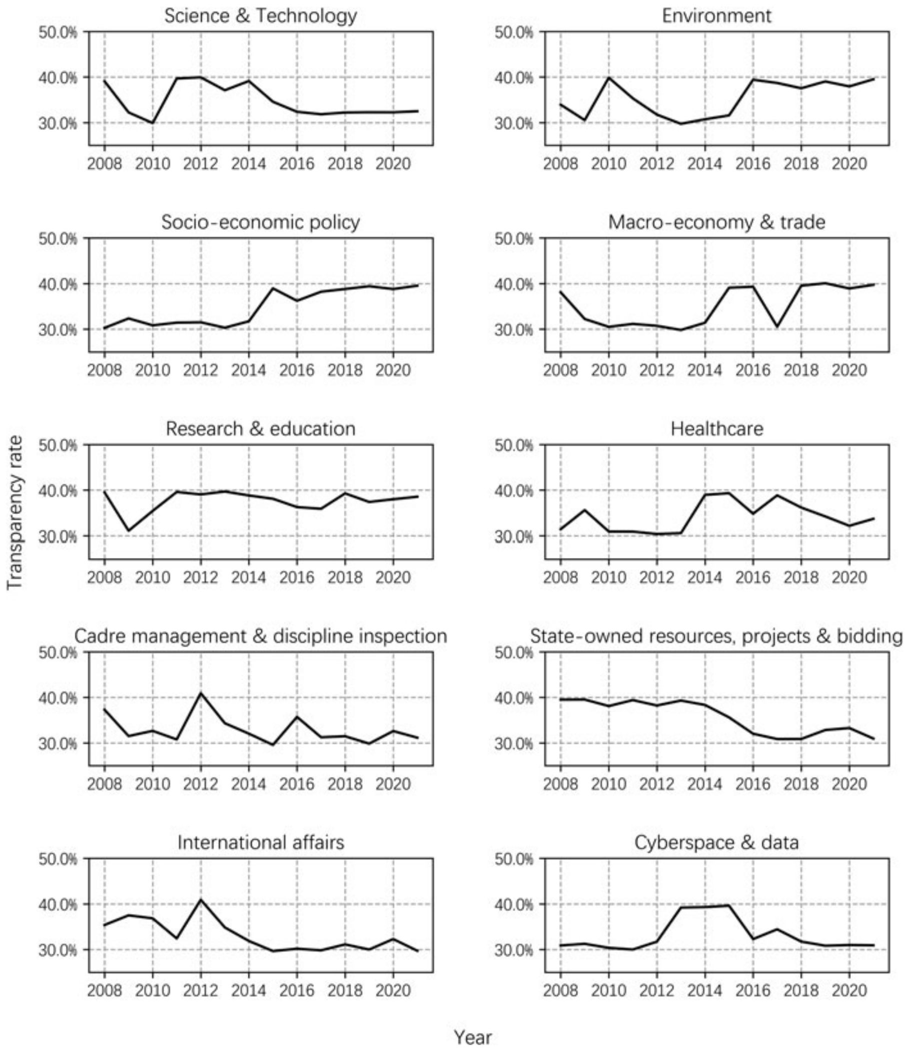

Figure 3. Transparency Rates by Topic, 2008–2021

were unavailable owing to issues uploading the websites.[43] The remainder of the links were inaccessible as the documents had actually disappeared.

For 7.7 per cent of all links, the websites had undergone updates to their infrastructure, causing the links to break. Technically, these documents could have been migrated elsewhere, but this is not always the case. One example is a website update by the Ministry of Housing and Urban-Rural Development between autumn 2021 and early 2022. Prior to the update, the website had hosted an extensive archive of local policy documents; however, the website managers did not transfer this archive to the new environment. This led to the disappearance of many documents, such as some of the initial local plans for the social credit system.[44]

Authorities may also delete information on the grounds that it is outdated and no longer relevant to ongoing policy. Policy documents referring to the OGI repeatedly emphasize "cleaning up"

[43] For a discussion of geo-blocking and other access challenges, see Brussee and Von Carnap 2024.

[44] For example, Chongqing People's Government 2003.

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

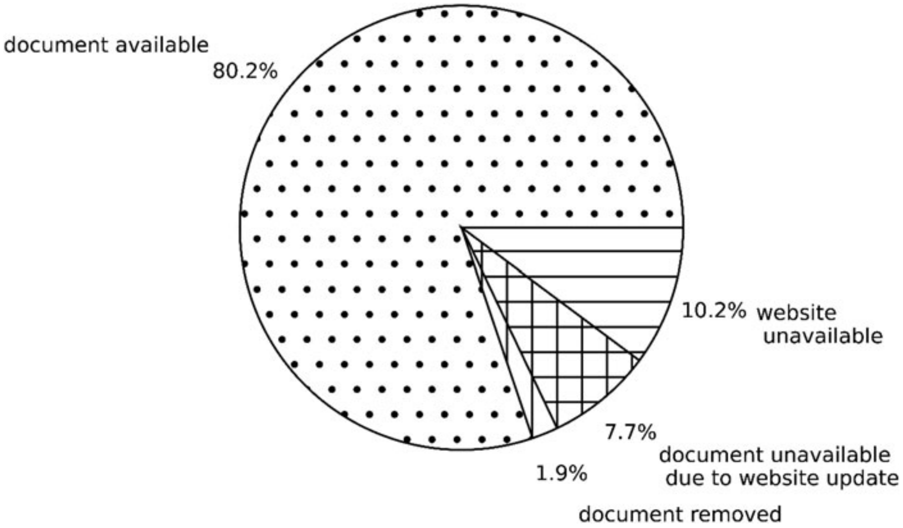

Figure 4. Availability of Links to Policy Documents Originally Retrieved in 2021, as of October 2023

outdated and expired content.[45] While PKULaw has a segment that hosts "cancelled" documents, it is unclear how comprehensive that segment is. Furthermore, it is certain that not all government websites do the same. Where they do not, such documents then disappear from the government websites altogether. Finally, government initiatives can become controversial after publication, which can lead to authorities cancelling the initiative and then attempting to erase all trace of it. For instance, following an online backlash, local authorities scrambled to take down documents that authorized the blacklisting of Chinese citizens who failed to get Covid-tested during the pandemic.[46] Both types of disappearance relate to information that is consciously deleted from a website and make up about 1.9 per cent of the total links tested here.

The focus on disappearing documents, however, invites a discussion of *appearing* documents, i.e. those that are released to the public a long time after the policy has been issued internally or come into effect. Figure 5 below displays the mean number of days from issuance of a document to its publication for seven government websites. In some cases, such as for the State Council and the Sichuan government's website, there used to be average delays of one to two years between issuance and publication. However, this delay has since become more standardized, at around 10 to 20 days. This indicates that while retroactive publication used to be a major source of variation in the earlier years of the OGI initiative, it is unlikely to significantly distort findings in more recent years that are based on very large datasets.[47] Nevertheless, retroactive publication following a long delay still occurs on a small scale. Henan's provincial government, for instance, delayed publication of four documents, which were originally issued in 2017, by two to three years. Hence, it remains important to reflect on the missingness problem in this domain, too, especially for studies using smaller subsets of data.

Mitigating Variation

As this paper demonstrates, there is significant variation in policy transparency and document availability over time. Transparency originally improved between 2008 and the mid-2010s. Today,

[45] See, e.g., People's Government of the Ningxia Autonomous Region 2023.

[46] Brussee 2023, Ch. 6–7.

[47] It does require researchers to carefully distinguish between the issuing date (*yinfa riqi*) and the publishing date (*gongbu riqi*) of a document, which are not always neatly indicated separately.

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

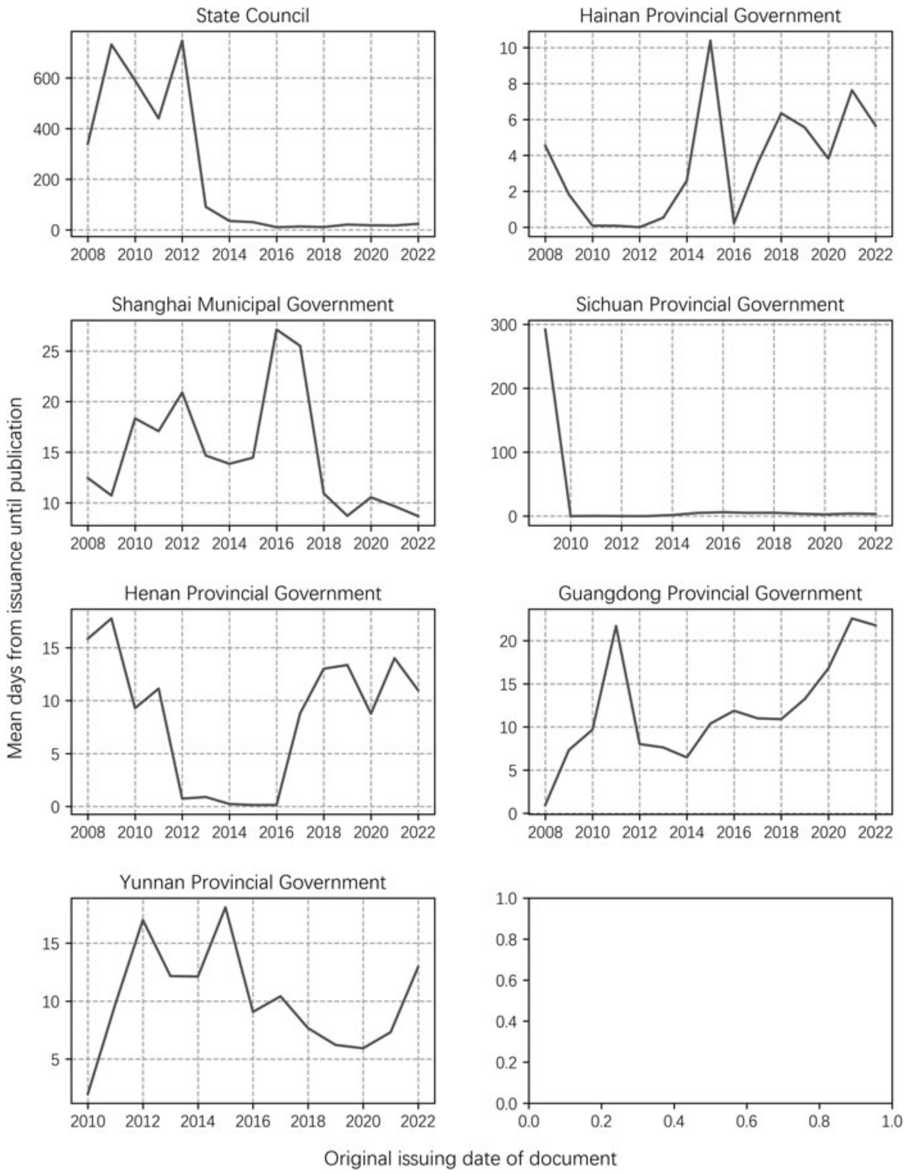

Figure 5. Mean Number of Days from Issuance of Document to Public Publication by Year of Original Issuance
Note: Data taken from seven government databases. Y-axis scales differ for better readability.

however, transparency is in decline in several fields, especially in fields where there are related geopolitical tensions. There is also significant variation among types of documents, with top-level policies seeing significantly higher disclosure rates than lower-level documents. Variation among topics appears primarily in the extent to which a topic is related to national security or citizens' daily lives. Finally, disappearance of documents is a real challenge for research. Thus, studies working with policy data must be open about how they mitigate missingness.

This paper's findings offer several guidelines. The low transparency rates for 2008 across the board indicate that any pre-2008 data are spotty at best; this was also the first year that the OGI regulations were implemented nationwide. Hence, there are fundamental questions surrounding the validity of

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

(quantitative) causal inferences based on policy texts that go further back than this date. For some localities, data are only somewhat representative starting from the mid to late 2010s.

While normalization is a key approach to mitigate missingness across time, variation is not randomly distributed: transparency has increased for certain topics yet decreased for others. Therefore, normalization alone is typically insufficient. Dealing with non-random variation can include controlling for policy type. By selecting only policy types for which disclosure is more standardized (for example, opinions, regulations and plans, instead of notices or reports), there are better chances that findings are not affected by external variation. Similarly, researchers might use this paper's findings on topical or local variation when selecting appropriate case studies. These strategies align with practices used for the study of court judgments, where scholars have recommended avoiding case types that suffer the greatest missingness and where officials have the greatest incentives for selective disclosure.[48] Another best practice is to combine quantitative inquiry with qualitative research.[49] Finally, the scholarly community needs to ensure sources remain available despite deletion or access challenges – for instance, by archiving sources through online tools or even creating entirely new archives that are hosted outside China.[50]

This paper's findings have broader implications. First, missingness can be indicative of internal government logics.[51] Thus, the findings in this paper double as a window into the internal government logics pertaining to policy transparency. More research can be done to add more depth to these findings and further leverage missingness in this and other fields. Second, missingness is not just an issue for policy documents; it affects virtually every study that relies on information sources curated by Chinese authorities. Many of the approaches developed here can also act as a basis for best practices in other fields.

This paper invites broader reflection on the fragility of our knowledge base and the use of convenient datasets in China studies. Policy documents are not propaganda, yet the fact that all these data are available to "us" also suggests that their availability serves a political purpose. The developments highlighted throughout this paper suggests that this curation of information sources is only likely to intensify. Understanding the context in which these sources are produced and what can – and, more importantly, cannot – be learned from them is crucial. While this paper focuses on variation and missingness, it is important to triangulate findings from policy documents (the paper reality) with actual lived experiences. More critical reflection on this is needed.

Acknowledgements. The author is grateful to Rogier Creemers and Florian Schneider for their feedback on earlier versions of this manuscript, as well as to the members of the "Puzzle Group" for their comments. This research is funded by the Dutch Research Council (NWO) under grant No. 406.22.CTW.013.

Competing interests. None.

References

Alves, Ana Cristina, and Su-Hyun Lee. 2022. "China's BRI developmental agency in its own words: a content analysis of key policy documents." *World Development* **150**, 105715.

Ang, Yuen Yuen. 2024. "Ambiguity and clarity in China's adaptive policy communication." *The China Quarterly* **257**, 20–37.

Batke, Jessica, Julia Breuer and Matthias Stepan. 2016. "Open government in China: bound to improve, within bounds." *The Asia Dialogue*, https://web.archive.org/web/20200924175415/https://theasiadialogue.com/2016/11/11/open-government-in-china-bound-to-improve-within-bounds/. Accessed 24 September 2020.

Brussee, Vincent. 2023. *Social Credit: The Warring States of China's Emerging Data Empire*. Singapore: Palgrave Macmillan.

Brussee, Vincent, and Kai Von Carnap. 2024. "The increasing challenge of obtaining information from Xi's China." *MERICS*, https://web.archive.org/web/20240229182619/https://merics.org/en/report/increasing-challenge-obtaining-information-xis-china. Accessed 29 February 2024.

[48] Xi 2022, 6.

[49] Ibid.; Heffer and Schubert 2023.

[50] This was also recommended in Tiffert 2019.

[51] Liebman et al. 2020.

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

Chen, Hui, Rongyu Xian, Hongqi Yu, Zheng Yang, Yaoyao Zhang, Duan Junjun, Jing Wang, et al. 2023. "Buweiji zhengfu menhu wangzhan zhengwu gongkai pingguo zhibiao tixi yanjiu" (Research on the evaluation index system of government affairs openness on the ministerial government portal). *Ziran ziyuan xinxihua* **4**, 10–17.

Chen, Lei, Zhuang Liu and Yingmao Tang. 2022. "Judicial transparency as judicial centralization: mass publicity of court decisions in China." *Journal of Contemporary China* **31**(137), 726–739.

Chongqing People's Government. 2003. "Guanyu yinfa Chongqing shehui xinyong tixi jianshe fang'an de tongzhi" (Notice on the issuance of Chongqing's social credit system constriction plan). On file with author.

Clark, George, Alex Gonye and Steven Miller. 2021. "Lessons from the German tank problem." *The Mathematical Intelligencer* **43**, 19–28.

Creemers, Rogier. 2021. "Party ideology and Chinese law." In Rogier Creemers and Susan Trevaskes (eds.), *Law and the Party in China: Ideology and Organisation*. Cambridge: Cambridge University Press, 31–64.

Cyrill, Melissa. 2018. "What is Made in China 2025 and why has it made the world so nervous?" *China Briefing*, 28 December, https://web.archive.org/web/20231011133816/https://www.china-briefing.com/news/made-in-china-2025-explained/. Accessed 11 October 2023.

deLisle, Jacques. 2017. "Law in the China model 2.0: legality, developmentalism and Leninism under Xi Jinping." *Journal of Contemporary China* **26**(103), 68–84.

General Office of the State Council. 2017. "2017 nian zhengwu gongkai gongzuo yaodian" (Work priorities for open government affairs, 2017), https://web.archive.org/web/20170323124634/http://www.gov.cn/zhengce/content/2017-03/23/content_5179996.htm. Accessed 23 March 2017.

General Office of the State Council. 2022. "2022 nian zhengwu gongkai gongzuo yaodian" (Work priorities for open government affairs, 2022), https://web.archive.org/web/20230630090021/https://www.gov.cn/gongbao/content/2022/content_5688781.htm. Accessed 30 June 2023.

Heffer, Abbey, and Gunter Schubert. 2023. "Policy experimentation under pressure in contemporary China." *The China Quarterly* **253**, 35–56.

Horsley, Jamie P. 2007. "Toward a more open China?" In Anne Florini (ed.), *The Right to Know: Transparency for an Open World*. New York: Columbia University Press, 54–91.

Horsley, Jamie P. 2019. "Transparency, accountability and access to information." In Sarah Biddulph and Joshua Rosenzweig (eds.), *Handbook on Human Rights in China*. Cheltenham: Edward Elgar, 516–544.

Hu, Huiming. 2023. "Liqing bianjie guifan liucheng – zhengfu xinxi gongkai tiaoli shijian nanti jiexi" (Clarifying the boundaries and standardizing the process: an analysis of the practical difficulties of the OGI regulations). *Zhengce jiedu* **12**, 17–19.

Huang, Cui, Shutao Wang, Jun Su and Peiqiang Zhao. 2020. "A social network analysis of changes in China's education policy information transmission system (1978–2013)." *Higher Education Policy* **33**, 323–345.

Liebman, Benjamin, Margaret Roberts, Rachel Stern and Alice Wang. 2020. "Mass digitization of Chinese court decisions: how to use text as data in the field of Chinese law." *Journal of Law and Courts* **8**(2), 177–201.

Liebman, Benjamin, Rachel Stern, Xiaohan Wu and Margaret Roberts. 2023. "Rolling back transparency in China's courts." *Columbia Law Review* **123**(8), 2407–82.

Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. "'Fightin' words: lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* **16**(4), 372–403.

Nan, Yan, Tieying Feng, Yuqun Hu and Xinzhu Qi. 2020. "Understanding aging policies in China: a bibliometric analysis of policy documents, 1978–2019." *International Journal of Environmental Research and Public Health* **17**(5956), 1–15.

People's Government of Jiangsu Province. 2021. "Shisiwu keji chuangxin guihua" (14th Five-Year Plan on Science, Technology and Innovation), https://archive.ph/8HgmY. Accessed 12 October 2023.

People's Government of Ningxia Autonomous Region. 2023. "2023 nian zhengwu gongkai gongzuo yaodian" (Work priorities for open government affairs, 2023), https://archive.ph/fxDE6. Accessed 27 September 2023.

Shambaugh, David. 2024. "The evolution of American contemporary China studies: coming full circle?" *Journal of Contemporary China* **33**(146), 314–331.

State Council. 2018. "Guowuyuan gongzuo guize" (Work regulations for the State Council), https://web.archive.org/web/20230517100800/http://www.gov.cn/zhengce/content/2018-07/02/content_5302908.htm. Accessed 17 May 2023.

State Council. 2023. "Guowuyuan gongzuo guize" (Work regulations for the State Council), https://web.archive.org/web/20230517114121/http://www.gov.cn/zhengce/content/2023-03/24/content_5748128.htm. Accessed 17 May 2023.

Stromseth, Jonathan, Edmund Malesky and Dimitar Gueorguiev. 2017. *China's Governance Puzzle: Enabling Transparency and Participation in a Single-party State*. Cambridge: Cambridge University Press.

Tiffert, Glenn D. 2019. "Peering down the memory hole: censorship, digitization, and the fragility of our knowledge base." *The American Historical Review* **124**(2), 550–568.

Van de Ven, Hans. 1995. "The emergence of the text-centered Party." In Tony Saich and Hans Van de Ven (eds.), *New Perspectives on the Chinese Revolution*. New York: Routledge, 5–32.

Wu, Guoguang. 1995. "'Documentary politics': hypotheses, process, and case studies." In Carol Lee Hamrin, Suisheng Zhao and A. Doak Barnett (eds.), *Decision-Making in Deng's China*. New York: Routledge, 24–38.

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

Wu, Xiaohan, Margaret Roberts, Rachel Stern, Benjamin Liebman, Amarnath Gupta and Luke Sanford. 2022. "Augmenting serialized bureaucratic data: the case of Chinese courts." 21st Century China Center Research Paper No. 2022-11, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4124433. Accessed 19 September 2023.

Xi, Chao. 2022. "How the Chinese judiciary works." *China Review* **22**(3), 1–8.

Zhang, Qiang, Qibin Lu, Deping Zhong and Xuanting Ye. 2018. "The pattern of policy change on disaster management in China: a bibliometric analysis of policy documents, 1949–2016." *International Journal of Disaster Risk Science* **9**, 55–73.

Zhou, Hanhua. 2016. "Dazao shengji ban zhengwu gongkai zhidu" (Creating an upgraded version of the government affairs publicity system). *Xingzheng faxue yanjiu* **3**, 3–13.

Appendix: Data and Methods

Data Sources

Data was scraped from over 80 different portals, using the "Open Government Affairs" (*zhengwu gongkai* 政务公开) sub-sections of official websites. These include:

Data type	Examples
Policy depositories: national level	State Council policy database, Supreme People's Court OGI portal
Policy depositories: ministerial level	OGI portals of all major ministries, including NDRC, MIIT, MOST, MEE, MOE, MOHURD, MCA, MOHRSS, MOA, NHC, PBOC, etc. Only peripheral ministries, such as the Ministry of Veteran Affairs, Water Management, etc. were not scraped.
Policy depositories: sub-ministerial units	OGI portals of a small selection of sub-ministerial units like the NEA.
Policy depositories: provincial level	OGI portals of all provincial governments.
Policy-adjacent sources	Xi Jinping Speeches Database, meeting records of State and Party organs including the State Council, Politburo Study Sessions, Central Commission for Comprehensively Deepening Reform, frontpage of *People's Daily*.

It should be noted that the dataset used for this paper is different from PKULaw, which is used in many of the studies cited throughout this paper. Unfortunately, PKULaw has restrictions on automatic retrieval of policies and, in addition, requires a licence to view the full content. This means large-scale analyses of its content are extremely difficult and it was not possible to replicate this paper's analysis to PKULaw. A brief review of data availability between the two data sources suggests that the differences are minor, i.e. in the 1–5% range. For instance:

Database	Period sampled	PKULaw	Scraped database	Difference
General Office of the State Council	2021	71	68	−4.4%
People's Government of Guangdong Province	All	4,566	4,678	2.4%
Ministry of Education	All	13,384	12,540	−6.7%

Some government websites have also started to implement restrictions on automated retrieval. For instance, the Ministry of Foreign Affairs Spokesperson database only allows retrieval of the last 1,000 results for any query.[52] This could create additional variation between data sources.

Distinctive Word Analysis

To conduct the distinctive word analysis, this paper relied on the Jieba software to automatically tokenize and segment words. Subsequently, it used Jieba to restrict results to only two-or-more character nouns, verbs and adjectives to conduct the analysis. As noted, it finally manually categorized and selected the keywords for display in the distinctive word analysis. This is for three reasons. First, not all keywords are informative. For instance, the keywords most strongly associated with public disclosure are "soliciting opinions" (*zhengqiu yijian* 征求意见). This is not particularly informative because this practice is, by its very nature, public. Other keywords that are not as informative include terms like "work" (*gongzuo* 工作), "issuance"

[52] https://www.mfa.gov.cn/web/wjdt_674879/fyrbt_674889/index.shtml

https://doi.org/10.1017/S0305741024000948 Published online by Cambridge University Press

(*yinfa* 引发), "to perfect" (*wanshan* 完善), etc. Second, some keywords are highly distinctive potentially because of data limitations. One highly distinctive word is "National Tax Administration" (*guojia shuiwu zongju* 国家税务总局), but this is most likely because this agency is not included in the web scraper. Third, keywords may be related to different functions. For instance, the keyword "opinion" (*yijian* 意见) specifically refers to the rubric of a document, not to policymaking on opinions.

In selecting keywords for the tables, I followed three guidelines:

1. The keyword must inform the reader about a clear topic or category that is associated with (non)transparency.
2. It must be verifiable that the keyword selected is not an artefact caused by limited data or by the word segmentation tools used.
3. There must be other, similar, keywords that show similar distinctiveness scores.

The full list of keywords and their distinctiveness scores are available in the GitHub repository of this project and can be independently verified.[53]

Dictionary Method

The dictionary method measures topics by the presence of keywords. For this paper, automated scripts coded each document according to whether its title contained one or more keywords related to a topic. In this way, a document could be coded with multiple topics. This is a logical approach, given that many documents are lengthy and can discuss many different topics within their contents. The table below provides examples of the keywords used to code each document. The full code and keyword lists can be found in the GitHub repository for the project.

Topic	Keyword examples
Science and technology	科技，技术，高新
Environment	环保, 环境, 废物
Socioeconomic policy	就业，养老，社会
Macroeconomy and trade	经济，贸易，价格
Research and education	教育，研究院，科研
Healthcare	卫生, 医, 药
Cadre management and discipline inspection	干部，检查，党内
State-owned resources, projects and bidding	采购，专项，国有
International affairs	国际，世界，外国
Cyberspace and data	网络，互联网，数据

Vincent BRUSSEE is a PhD candidate at Leiden University. He specializes in the application of natural language processing for contemporary Chinese policy analysis and is the author of *Social Credit: The Warring States of China's Emerging Data Empire* (2023, Palgrave Macmillan). Previously, he was an analyst at the Mercator Institute for China Studies (MERICS) in Berlin.

[53] https://github.com/zongtihuoguoguan/Policy-Transparency-China-2024

Cite this article: Brussee V (2024). Mitigating Missingness in Analysing Chinese Policy and Implications for the Fragility of Our Knowledge Base. *The China Quarterly* 1–15. https://doi.org/10.1017/S0305741024000948